Estatística não paramétrica

Professor: Dr. Pedro M. Almeida-Junior

11 de julho de 2021

Departamento de Estatística (UEPB)

Ementa

Unidade I

- Introdução a estatística não paramétrica
- Revisão dos conceitos de inferência
- Teste para uma amostra
  - Binomial
  - Qui-Quadrado
  - Teste Exato de Fisher
  - Teste de Kolmogorov-Sminorv
- **Testes duas amostras independentes:**
  - Teste da mediana
  - Teste U de Mann-Whitney
  - Teste de Kolmogorov Smirnov

Ementa

Unidade II

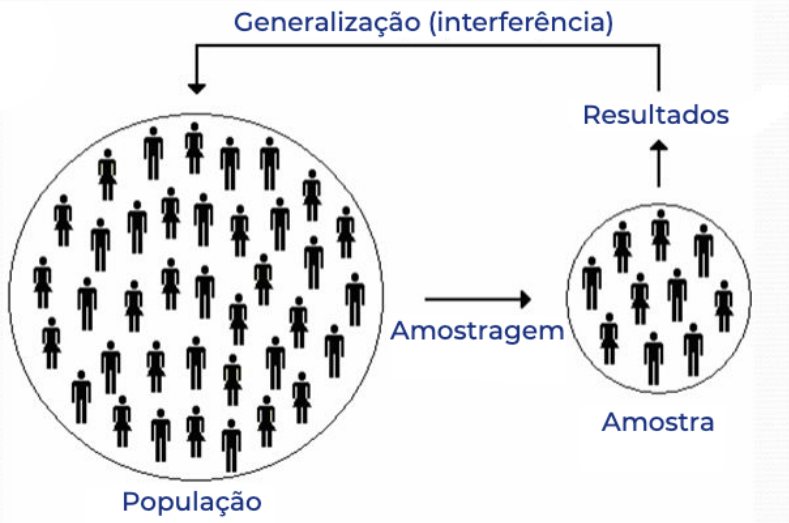
- Testes amostras pareadas:
  - Teste do Sinal
  - Teste de McNemar
  - Wilcoxon
- Testes de correlação:
  - Mann-Kendal
  - Spearman
- Testes para várias amostras (ANOVA):
  - Kruskall-Wallis
- Modelos não paramétricos

Revisão Inferência estatística

Inferência estatística

- Um tópico central da estatística é a inferência estatística.

- Na inferência estatística, procuramos tirar conclusões sobre um grande número de eventos com base na observação de apenas uma parte deles;



- A inferência estatística aborda dois tipos de problemas fundamentais
  - Estimação de parâmetros(Pontual e Intervalar);
  - Testes de hipóteses.

Inferência estatística

- Um problema comum em inferência estatística consiste em determinar, em termos de probabilidades, se as diferenças observadas entre duas amostras significam que sejam realmente diferentes entre si.

- Os processos de inferência permite determinar se a diferença pertence em um intervalo de valores que possam ser atribuídos ao acaso ou fatores aleatórios.

- Podemos fazer isso através de testes de hipóteses (ou intervalo de confiança) que podem ser classificados em paramétricos e não paramétricos.

Técnicas paramétricas e não paramétricas

- Supõem hipóteses sobre a natureza da população da qual foram extraídos os dados (como os valores relacionados a essa população são parâmetros, essas técnicas são conhecidas como paramétricas.);

- Essas técnicas baseam-se na hipótese de que os dados extraídos de uma população, cuja distribuição dos dados sejam conhecidas (distribuição normal) ou que dois conjuntos de dados tenham sido extraídos de populações com mesma variância

- Entretanto, existem técnicas que não exigem hipóteses sobre os parâmetros. Denomina-se técnicas não-paramétricas.

Estatística não-paramétrica

- A Estatística Não-Paramétrica pode ser definida como uma coleção de métodos estatísticos aplicada a conjuntos de dados onde as suposições distribucionais necessárias para aplicação de uma técnica clássica (Intervalo de Confiança, Teste de Hipótese) não são satisfatoriamente atendidas.

- Atualmente esta área da Estatística está bastante desenvolvida e os tópicos modernos são Estimação de Densidade, Regressão Não-Paramétrica e Semi-Paramétrica (Principais aplicações: *Machine Learning*).

- Algumas técnicas não-paramétricas são testes de postos ou testes de ordenaç e estas formas de identificação sugerem outro aspecto em que testes não-paramétricos diferem de testes paramétricos.

Enquanto um teste paramétrico pode focalizar sobre a diferença entre as médias de duas populações, o teste não-paramétrico análogo pode focalizar sobre a diferença das medianas.

Vantagens e Desvantagens (Estatística não-paramétrica)

VANTAGENS:

- Típicamente fazem menos suposições sobre os dados
- Podem ser utilizados para tratar dados que são simplesmente classificatórios ou categóricos
- São testes mais simples;
- São mais eficientes que os paramétricos quando não existe Normalidade.

DESVANTAGENS:

- Proporcionam um desperdício de informações, já que em geral não consideram a magnitude dos dados;
- Quando as suposições do modelo estatístico são atendidas, as técnicas não-paramétricas são menos eficientes que os paramétricos;

Testes de hipóteses

- Para chegarmos a uma decisão objetiva sobre uma hipótese particular, devemos dispor de um processo objetivo que nos permita rejeitar ou não rejeitar uma hipótese;

- O procedimento usualmente seguido envolve vários passos. São eles:
  1. Definir a hipótese nula (H_0)
  2. Escolher um teste de hipótese adequado ao tipo de dados para provar H_0.
  3. Especificar qual o nível de significância (α)
  4. Determinar a distribuição amostral da estatística do teste sob a hipótese nula
  5. Com base nos três últimos item, definir a região crítica (ou de rejeição)
  6. Calcular o valor da estatística do teste utilizando os dados obtidos das amostras. Rejeitar ou não a hipótese nula de acordo com a região crítica.

Hipótese Nula (H_0)

- A hipótese nula (H_0) é uma hipótese de "não-efeito" e é usualmente formulada com o propósito de ser rejeitada; ou seja, é a negação do ponto que se está tentando confirmar.

- Se ela é rejeitada, a hipótese alternativa (H_1) é confirmada.

- A hipótese alternativa é a afirmação operacional da hipótese de pesquisa do investigador.

- Quando queremos tomar uma decisão sobre diferenças, testamos H_0 contra H_1. Se H_0 não é rejeitada temos evidências em favor desta hipótese.

Exemplos de hipóteses

- Hipóteses para testar diferença entre **médias**:
  1. $H_0 : \mu_1 = \mu_2 \times H_1 : \mu_1 \neq \mu_2$
  2. $H_0 : \mu_1 \leq \mu_2 \times H_1 : \mu_1 > \mu_2$
  3. $H_0 : \mu_1 \geq \mu_2 \times H_1 : \mu_1 < \mu_2$

- Hipóteses para testar diferença entre **proporções**:
  1. $H_0 : p_1 = p_2 \times H_1 : p_1 \neq p_2$
  2. $H_0 : p_1 \leq p_2 \times H_1 : p_1 > p_2$
  3. $H_0 : p_1 \geq p_2 \times H_1 : p_1 < p_2$

Nível de significância e tamanho da amostra

- Quando a hipótese nula e a hipótese alternativa já foram estabelecidas e quando o teste estatístico apropriado já foi selecionado, o próximo passo é especificar um **nível de significância** (α) e selecionar um **tamanho para a amostra** (N)

- Nosso procedimento é rejeitar H_0 em favor de H_1 se um teste estatístico dá um valor cuja probabilidade associada de ocorrência sob H_0 é menor ou igual a alguma pequena probabilidade, usualmente denotada por α. **Esta probabilidade é chamada de nível de significância**.

- Valores comuns para α são 0,01, 0,05 e 0,10.

- se a probabilidade associada à ocorrência sob H_0 (por exemplo, quando a hipótese nula é verdadeira) de um valor particular fornecido por um teste estatístico (e mais valores extremos) é menor ou igual a α, rejeitamos H_0 em favor de H_1, a afirmação operacional da hipótese de pesquisa.

- Pode ser visto, então, que α dá a probabilidade de falsamente rejeitar H_0.

- Como a probabilidade α entra no processo de determinação de aceitação ou de rejeição de H_0, a necessidade de objetividade exige que α seja especificado antes que os dados sejam coletados

Erros Tipo I e II

- Existem dois tipos de erros que podem ser cometidos ao chegar a uma decisão sobre H_0:
  1. **Erro do Tipo I:** Rejeição da hipótese H_0 quando ela é, de fato, verdadeira.
  2. **Erro do Tipo II:** Não rejeitar a hipótese nula H_0, quando, de fato, ela é falsa.

- A probabilidade de cometer um erro do Tipo I é denotado por α.

$$P[\text{Erro Tipo I}] = \alpha$$

- A probabilidade de cometer o erro do Tipo II é usualmente denotado por β. Isto é,

$$P[\text{Erro Tipo II}] = \beta$$

Decisão	Situação Real	
	H_0 é verdadeira	H_0 é falsa
H_0 não é rejeitada	Decisão correta	Erro tipo II
H_0 é rejeitada	Erro tipo I	Decisão correta

O ideal seria que ambas as probabilidades de erro, α e β, fossem tão pequenas quanto possível, mas infelizmente, se o tamanho n da amostra está fixado, uma diminuição de β traz como consequência um aumento em α e vice-versa.

Controle do Erro tipo I

Por que Controlamos o Erro Tipo I ao invés do Erro Tipo II ?

Por definição, α é calculada supondo-se H_0 verdadeira, ou seja, a partir da distribuição de probabilidade única postulada por H_0. Por outro lado, β é calculada supondo H_1 verdadeira. Então, em geral há infinitos valores possíveis para β, correspondentes a infinitas distribuições de probabilidade.

Poder do teste

- Na prática é mais comum α e N serem especificados primeiramente. Uma vez que α e N tenham sido especificados, β é determinado.

- O poder de um teste é definido como a probabilidade de rejeitar H_0 quando ela é, de fato, falsa. Isto é,

$$\text{Poder} = 1 - P[\text{Erro Tipo II}] = 1 - \beta \tag{1}$$

Distribuição amostral

- Após um pesquisador ter escolhido um certo teste estatístico para usar com um conjunto de dados, a distribuição da estatística do teste precisa ser determinada.

- A **distribuição amostral** é a distribuição, quando H_0 é verdadeira, de todos os possíveis valores que alguma estatística (digamos, a média amostral $\bar{x}$) pode tomar quando a estatística é calculada a partir de muitas amostras de mesmo tamanho, extraídas da mesma população.

- **A probabilidade associada com a ocorrência sob** H_0 é a probabilidade de ocorrência sob H_0 de um valor tão extremo quanto o valor particular da estatística do teste.

Estatística dos testes

- Teoremas matemáticos são usados para assumir uma distribuição amostral para nossa estatística do teste. Estes teoremas, invariavelmente, envolvem suposições, e ao aplicar os teoremas precisamos manter em mente estas suposições.

- Usualmente tais suposições referem-se à distribuição da população e/ou ao tamanho da amostra. Um exemplo de um tal teorema é o **Teorema Central do Limite**.

- Quando uma variável é normalmente distribuída, sua distribuição é completamente caracterizada por sua média e por seu desvio-padrão.

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0, 1)$$

Região de Rejeição

- A distribuição amostral inclui todos os possíveis valores que a estatística do teste pode assumir.

- A região de rejeição consiste de um subconjunto destes valores possíveis e é escolhida de modo que a probabilidade sob H_0 de que a estatística do teste assuma um valor naquele subconjunto seja α.

- Em outras palavras, a região de rejeição consiste em um conjunto de possíveis valores que sejam tão extremos que, quando H_0 é verdadeira, a probabilidade de que a amostra no qual observamos forneça um valor entre eles seja realmente pequena (isto é, igual a α).

- A probabilidade associada à qualquer valor na região de rejeição é menor ou igual a α.

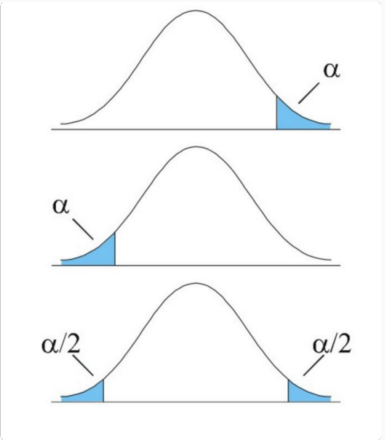
Região de Rejeição

- O tamanho da região de rejeição é expresso por α, o nível de significância. Se $\alpha = 0,05$, então o tamanho da região de rejeição compreende 5% da área total incluída sob a "curva" da distribuição amostral.

Teste Unilateral à direita

Teste Unilateral à esquerda

Teste Bilateral



Decisão

- Se o teste estatístico fornece um valor que está na região de rejeição, rejaitamos H_0.

- Se a probabilidade associada com a ocorrência, sob H_0, de um valor particular na distribuição amostral é muito pequena, podemos decidir que a hipótese nula é falsa

- Quando a probabilidade associada com um valor observado da estatística do teste é igual ou menor do que o valor previamente determinado de α, concluímos que H_0 é falsa. Tal valor observado é dito **significante**.

ESCOLHA DO TESTE ESTATÍSTICO ADEQUADO

- É importante a definição de critérios que nos ajudem a decidir qual o teste ideal para determinado problema.

- Poder do Teste $(1 - \beta)$: O teste que apresenta uma maior probabilidade de rejeitar H_0 quando H_0 é falsa, entre todos os testes de nível α deve ser escolhido. Mas só isto não basta e nem sempre é simples de ser obtido, portanto precisamos de outras informações para escolher o teste mais adequado:
  - Como foi obtida a amostra, ou seja, o plano experimental;
  - Natureza da População (pessoas, objetos, áreas, animais, etc.);
  - Tipo de Mensuração dos dados (escala de mensuração).

- Quando se usa um teste paramétrico existe uma série de pressupostos a serem verificados, além do nível mínimo de mensuração exigido ser a escala intervalar. No caso não-paramétrico, o primeiro critério a ser verificado deve ser o nível de mensuração dos dados.

NÍVEL DE MENSURAÇÃO

Modelo Estatístico

- A cada teste estatístico existe um modelo e uma exigência da mensuração. O teste é válido sob certas condições, e o modelo e a exigência da mensuração especificam estas condições.

- Precisamos examinar a situação e determinar se é razoável ou não assumir que o modelo é correto.

- Será que o teste que vou assumir serve para o tipo de variável que estou interessado em estudar ? Aqui discutiremos quatro níveis ou tipos de mensuração – nominal, ordinal e intervalar. As implicações de cada um para a interpretação de testes estatísticos.

Escala nominal (ou Categórica)

- É o mais baixo nível de mensuração. Utiliza símbolos ou números simplesmente para distinguir elementos em diferentes categorias (como um nome), não havendo entre eles, geralmente, possibilidade de comparação do tipo maior- menor ou melhor-pior).

  Exemplos:

  - Masculino (M), Feminino (F)
  - Objeto não defeituoso (1), Objeto defeituoso (0)
  - Casado(1), solteiro(2), divorciado(3) e viúvo(4).

- **Operações Admissíveis:** Os únicos tipos de estatísticas descritivas admissíveis são aqueles que não seriam alterados por alguma transformação – Por exemplo a moda e contagem da frequência.

Escala Ordinal

- Utiliza números apenas para classificarmos elementos numa ordem crescente ou decrescente. Existe assim algum tipo de relação entre as categorias embora a diferença entre elas seja de difícil quantificação.

  Exemplos:

  - Classes sócio econômicas: (Baixa, média, alta)
  - Patentes do Exército (soldado, cabo, sargento, etc)
  - Opinião de um determinado produto (Ruim, Regular, Bom, Muito bom, Excelente)

- **Operações Admissíveis:** uma transformação que não muda a ordem das classes é completamente admissível. A estatística mais apropriada para descrever a tendência central dos escores em uma escala ordinal é a mediana, pois, relativamente à distribuição dos escores, a mediana não é afetada por mudanças em quaisquer escores que estão acima ou abaixo dela.

Escala Intervalar

- Ocorre quando a escala tem as características da escala ordinal e ainda é possível quantificar a diferença (distância) entre dois números desta escala.
  Exemplos:
  - Temperatura
  - Peso
  - Altura
- **Operações Admissíveis:** Todas as estatísticas paramétricas comuns (médias, desvios-padrão, correlações, etc.) são aplicáveis aos dados em uma escala intervalar.

RESUMO

- A natureza dos dados nos indicam qual melhor teste a ser utilizado
- As suposições dos testes paramétricos devem ser checadas. Caso existam violações das suposições: Aplica-se a abordagem não paramétrica

Teste para Médias de uma amostra

Teste sobre a Média de uma Normal com Variância Desconhecida

- Baseado na distribuição t de Student, vamos testar hipóteses sobre μ, quando a variância populacional é desconhecida.
- Este Teste é denominado de teste t de Student
- As hipóteses nula e alternativa do teste são dadas a seguir

$$\begin{aligned} \mathrm{H}_0 &: \mu = \mu_0 \\ \mathrm{H}_1 &: \mu \neq \mu_0 \end{aligned}$$

A hipótese alternativa poderia ser $\mu > \mu_0$ ou $\mu < \mu_0$, o que mudaria apenas a região de rejeição de bilateral para unilateral (à direita ou à esquerda)

- A estatística do teste sob a hipótese nula é

$$T = \frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{S} \sim t_{(n-1)}$$

em que n é o tamanho da amostra, $\bar{X}$ é a média amostral, S é o desvio amostral e μ_0 é a média quando assumimos H_0 é verdadeira e $t_{(n-1)}$ representa uma distribuição t de Student com $n-1$ graus de liberdade.

- Logo, precisamos obter o valor crítico (t_c) baseado na distribuição t de Student, por exemplo para um teste bilateral

$$P(-t_c \leq T \leq t_c) = \alpha$$

ou unilateral,

$$P(T \geq t_c) = \alpha$$

Exemplo

Exemplo: Um fabricante afirma que seus cigarros contém não mais que $30\mathrm{mg}$ de nicotina. Uma amostra de 25 cigarros fornece média de 31,5 mg e desvio padrão de $3\mathrm{mg}$. No nível de 5%, os dados refutam ou não a afirmação do fabricante?

Teste para Médias de duas amostras independentes

Testes para duas amostras independentes

- A qui temos dados na forma de duas amostras, extraídas independentemente de cada população.

- É muito comum em experimentos do tipo "controle"versus "tratamento", nos quais o interesse principal é verificar o efeito desse último.

- 0 caso típico é aquele de comparar uma nova droga com uma padrão, usadas para o tratamento de uma doença.

Exemplos de Aplicações

1. Um curso de Estatística é ministrado pela televisão para um grupo de alunos e ao vivo para outro grupo. Podemos testar a hipótese de que o curso ao vivo é mais eficaz que o curso por meio da televisão.

2. Podemos comparar o efeito de duas rações, A e B, sobre o crescimento de porcos. Dois grupos de porcos em crescimento foram alimentados com as duas rações e após cinco semanas verificam-se quais foram os ganhos de peso dos porcos dos dois grupos.

3. 20 canteiros foram plantados com uma variedade de milho. Em dez deles um novo tipo de fertilizante é aplicado e nos outros um fertilizante padrão. Examinando-se as produções dos dois canteiros, queremos saber se há diferenças significativas entre as produções.

TESTES PARA COMPARAÇÃO DE DUAS MÉDIAS

- Temos basicamente duas categorias de problemas que se enquadram nessa situação: os testes pareados e os não-pareados.

- O planejamento do experimento é diferente nas duas categorias. No caso dos testes não-pareados, temos duas populações distintas e uma determinada característica (uma variável quantitativa) de modo que a cada elemento de ambas as populações pode ser associado o valor dessa variável que lhe corresponde.

- Já para os testes pareados, os dados são fornecidos aos pares, e a idéia é comparar diretamente, entre si, os dois elementos de cada par. Por exemplo, comparar o efeito de determinado tratamento, antes e depois.

Teste t para duas amostras independentes

- Aqui temos duas populações distintas e uma determinada característica, que será denotada por X em uma população e por Y na outra. Queremos comparar as médias populacionais de X e de Y.

- Para isso são coletadas amostras, uma em cada população, e, para cada elemento dessas amostras, mede-se o valor da característica considerada.

- Para usar o teste t , assumimos algumas suposições:
  1. As variáveis aleatórias X e Y são normais, ou sejam, nossos dados são normais.
  2. As amostras são independentes
  3. Os desvios-padrão populacionais σ_X e σ_Y são iguais, isto é, $\sigma_X = \sigma_Y = \sigma$, embora o seu valor σ seja desconhecido

Hipóteses do teste t

- Queremos testar a hipótese nula de médias populacionais iguais contra a alternativa de médias populacionais diferentes (maior ou menor)

- Isto é,

$$\mathrm{H}_0 : \mu_\mathrm{X} = \mu_\mathrm{Y} \text{ contra } \mathrm{H}_1 : \mu_\mathrm{X} \neq \mu_\mathrm{Y}$$
$$\mathrm{H}_0 : \mu_\mathrm{X} = \mu_\mathrm{Y} \text{ contra } \mathrm{H}_1 : \mu_\mathrm{X} > \mu_\mathrm{Y}$$
$$\mathrm{H}_0 : \mu_\mathrm{X} = \mu_\mathrm{Y} \text{ contra } \mathrm{H}_1 : \mu_\mathrm{X} < \mu_\mathrm{Y}$$

- A estatística de teste a ser usada é

$$\mathrm{T} = \frac{\overline{\mathrm{X}} - \overline{\mathrm{Y}}}{\sqrt{S_{\mathrm{p}}^2 \left(\frac{1}{\mathrm{m}} + \frac{1}{\mathrm{n}}\right)}}$$

em que:
$\bar{X}$ é a média amostral dos $\mathrm{X}_{\mathrm{i}}'s, \mathrm{i} = 1, \ldots, \mathrm{m}$
$\bar{Y}$ é a média amostral dos $\mathrm{Y}_{\mathrm{i}}'s, \mathrm{j} = 1, \ldots, \mathrm{n}$
S_x^2 é a variância amostral dos $X_i's, i = 1, \ldots, m$
S_y^2 é a variância amostral dos $\mathrm{Y}_j's, \mathrm{j} = 1, \ldots, \mathrm{n}$
$S_p^2 = \frac{(m-1)S_x^2 + (n-1)S_Y^2}{m+n-2}$ é o estimador combinado de σ^2 .

Distribuição amostral e decisão do teste t

Prova-se que, se H_0 é verdadeira, a variável aleatória T segue uma distribuição t de Student com $\mathrm{m} + \mathrm{n} - 2$ graus de liberdade. Assim, uma vez escolhido o nível de significância α, rejeita-se H_0 se $|\mathrm{T}_{\mathrm{obs}}| > \mathrm{t}_{1-\frac{\alpha}{2}}$.

Exemplo

Exemplo: Duas técnicas de venda são aplicadas por dois grupos de vendedores: a técnica A, por 12 vendedores, e a técnica B, por 15 vendedores. Espera-se que a técnica B produza melhores resultados. No final de um mês, obtiveram-se os resultados na Tabela seguinte.

Tabela 1: Dados para duas técnicas de vendas

Dados	Vendas	
	Técnica A	Técnica B
Média	68	76
Variância	50	75
Vendedores (n)	12	15

Vamos testar, para o nível de significância de 5%, se há diferenças significativas entre as vendas resultantes das duas técnicas. Informações adicionais permitem supor que as vendas sejam normalmente distribuídas, com uma variância comum σ^2, desconhecida.

Testes não-paramétricos que vamos ver no curso

1. Testes para uma amostra (Aderência)
2. Testes para duas amostras (independentes e dependentes)
3. Análise de variância (independentes e dependentes)
4. Medidas de associação e testes de significância